Diversity-Preserving K-Armed Bandits, Revisited

Hédi Hadiji, Sébastien Gerchinovitz, Jean-Michel Loubes, Gilles Stoltz

Diversity-Preserving K–Armed Bandits, Revisited

Hédi Hadiji HEDI.HADIJI@MATH.U-PSUD.FR
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, Orsay, France

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@IRT-SAINTEXUPERY.COM
Institut de recherche technologique Saint Exupéry, Toulouse
Institut de mathématiques de Toulouse, Université Paul Sabatier, Toulouse

Jean-Michel Loubes JEAN-MICHEL.LOUBES@MATH.UNIV-TOULOUSE.FR
Institut de mathématiques de Toulouse, Université Paul Sabatier, Toulouse

Gilles Stoltz GILLES.STOLTZ@MATH.U-PSUD.FR
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, Orsay, France

Editor: Under Review

Abstract

We consider the bandit-based framework for diversity-preserving recommendations introduced by Celis et al. (2019), who approached it mainly by a reduction to the setting of linear bandits. We design a UCB algorithm using the specific structure of the setting and show that it enjoys a bounded distribution-dependent regret in the natural cases when the optimal mixed actions put some probability mass on all actions (i.e., when diversity is desirable). Simulations illustrate this fact. We also provide regret lower bounds and briefly discuss distribution-free regret bounds.

Keywords: Multi-armed bandits; UCB strategy; diversity; regret bounds: upper and lower bounds

1. Setting and literature review

We consider stochastic bandit models with finitely many arms. All of them are desirable actions, though some lead to higher payoffs. Effective (regret-minimizing) algorithms are bound to play the optimal arm(s) an overwhelming fraction of time. Celis et al. (2019) refer to this effect as polarization and introduce a model to avoid it. We suggest the alternative terminology of preserving diversity. A general formulation of the diversity-preserving bandit model is provided below and is summarized in Protocol 1. Our aim in this article is to deepen and improve on the results obtained by the mentioned reference; see Sections 1.2 and 1.3 for details.

Extended literature review. A general discussion of the notions of diversity and fairness in stochastic and adversarial bandits is provided in Appendix B.

Diversity-preserving bandits As in traditional K–armed bandits, K probability distributions $\nu_1, \dots, \nu_K$ associated with each arm are considered, with expectations denoted by $\mu_1, \dots, \mu_K$. These distributions are unknown to the learner but belong to a known set of possible distributions, called a model $\mathcal{D}$. In this article we consider mainly the bandit model $\mathcal{D}_{[0,1]}$ of all probability measures supported on $[0,1]$, that is, we assume that rewards can be distributed according to any distribution bounded in $[0,1]$. An exception to this is the lower bound in Section 3, which we formulate on a generic model $\mathcal{D}$ satisfying a mild assumption.

The learning protocol is the following. An arm $A_t \in [K]$ is picked among K choices at each round, where we denote by $[K]$ the set $\{1, \dots, K\}$. The learner then obtains a payoff Y_t drawn at

random according to ν_{A_t} given that choice. This is the only observation made (the learner does not know what it would have obtained with a different choice). However, the distinguishing feature of the bandit model by Celis et al. (2019) is that the choice of A_t is made in two steps, as follows. First, a distribution $\underline{p}_t$ over the arms is picked, in some known closed set $\mathcal{P}$, which quantifies diversity (specific examples are given below); then, the arm A_t is drawn at random according to $\underline{p}_t$. Following game-theoretic terminology, we will call $a \in [K]$ pure actions or ams, and $\underline{p} \in \mathcal{P}$ mixed actions or probabilities.

We measure performance in terms of expected payoffs. The expected payoff at round t may be computed by repeated applications of the tower rule:

$$\mathbb{E}[Y_t \mid A_t, \underline{p}_t] = \mu_{A_t}\,, \quad \text{thus} \quad \mathbb{E}\big[Y_t \mid \underline{p}_t\big] = \sum_{k\in[K]} p_{t,k}\,\mu_k \overset{\text{def}}{=} \langle \underline{p}_t, \underline{\mu}\rangle\,, \quad \text{thus} \quad \mathbb{E}[Y_t] = \mathbb{E}\big[\langle \underline{p}_t, \underline{\mu}\rangle\big]\,.$$

Maximizing the cumulative expected payoff of a policy amounts to minimizing the expected regret defined as

$$R_T = T \max_{\underline{p}\in\mathcal{P}} \langle \underline{p}, \underline{\mu}\rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle \underline{p}_t, \underline{\mu}\rangle\right].$$

In the definition of the regret, the comparison is made with respect to the expected payoff that would have been obtained by picking at each round a best diversity-preserving distribution over the arms.

Protocol 1 summarizes the setting and aim.

Protocol 1 Diversity-preserving stochastic bandits (Celis et al., 2019)

Known parameters
Arms $1, \ldots, K$ and model $\mathcal{D}$ of distributions for the arms
Closed set $\mathcal{P}$ of diverse enough probability distributions over the arms
Unknown parameters
Probability distributions $\underline{\nu} = (\nu_1, \ldots, \nu_K)$ in $\mathcal{D}$, with expectations $\underline{\mu} = (\mu_1, \ldots, \mu_K)$

for $t = 1, 2, \ldots$ **do**
Pick a distribution $\underline{p}_t = (p_{t,1}, \ldots, p_{t,K}) \in \mathcal{P}$ over the arms
Draw at random an arm $A_t \sim \underline{p}_t$
Get and observe a payoff $Y_t \sim \nu_{A_t}$ drawn at random according to ν_{A_t} given A_t
end for

Aim
Minimize the expected regret $\quad R_T = T \max_{\underline{p}\in\mathcal{P}} \langle \underline{p}, \underline{\mu}\rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle \underline{p}_t, \underline{\mu}\rangle\right]$

1.1. Examples of diversity-preserving sets $\mathcal{P}$ of distributions over the arms

Simplest example. The simplest requirement is that each arm should be played with some minimal probability $\ell > 0$, which corresponds to $\mathcal{P} = \big\{\underline{p} : \ \forall a \in [K],\ \ p_a \geqslant \ell\big\}$. This constraint makes sense in online advertisement: all offers need to be displayed a significant fraction of the time and get a significant chance to be selected.

More generally, Celis et al. (2019) indicate that one could group arms intro groups $G_1, \ldots, G_N$ of similar arms and impose minimal probabilities $\ell_1, \ldots, \ell_N > 0$ as well as maximal probabilities $u_1, \ldots, u_N < 1$ for each group defined as:

$$\mathcal{P} = \left\{ \underline{p} : \ \forall g \in [N], \ \sum_{a \in G_g} p_a \in [\ell_g, u_g] \right\}.$$

Note that the sets $\mathcal{P}$ considered above are finite polytopes[1].

Maintaining a budget. The last example can be generalized as follows: every pure action a is associated with N costs $c_a^{(1)}, \dots, c_a^{(N)}$ in $\mathbb{R}$, accounting for limited resources or environmental costs like the amount of carbon emissions generated from taking the action. The model can handle negative costs (e.g., negative carbon emissions). When a player picks a pure action A_t according to the mixed action $\underline{p} = (p_1, \dots, p_K)$, the N expected costs associated with her choice are

$$\sum_{a \in [K]} p_a c_a^{(1)}, \ \dots, \ \sum_{a \in [K]} p_a c_a^{(N)} .$$

In this case, a reasonable objective for the player is to maximize her payoff under the constraints that, for all $n \in [N]$, the n–th expected cost of her actions be kept under a certain level $u_n \in \mathbb{R}$. This amounts to playing under Protocol 1 with the probability set

$$\mathcal{P} = \left\{ \underline{p} : \ \forall n \in [N], \ \sum_{a=1}^{K} p_a c_a^{(n)} \leqslant u_n \right\}.$$

This set is again a finite polytope. Note that the name "diversity-preserving" was inspired by the example of the previous paragraph and is perhaps less relevant in the present example.

1.2. Algorithms considered and regret bounds achieved by Celis et al. (2019)

Celis et al. (2019) approach the setting considered above by seeing it a special case of linear stochastic bandits. Indeed, we showed that the expected payoff obtained at each round equals $\mathbb{E}\big[Y_t \mid \underline{p}_t\big] = \big\langle \underline{p}_t, \underline{\mu} \big\rangle$, which is a linear function of the probability $\underline{p}_t$ picked. This observation opens the toolbox of algorithms to deal with stochastic linear bandits (with action set $\mathcal{A} = \mathcal{P}$, see Appendix C for more details) to solve the considered problem; this is exactly what Celis et al. (2019) do. For instance, they use the LinUCB (linear upper confidence bound), also known as OFUL (optimism in the face of uncertainty for linear bandits), strategies introduced by Li et al. (2010) and Chu et al. (2011) and further studied by Abbasi-Yadkori et al. (2011). They obtain regret bounds of order at best $K(\ln T)^2/\Delta$, with the notation of Theorem 1, which is of the same order as our most general bound (Case 1 of Theorem 1). However, their algorithm is less computationally efficient, and the lower order terms, as well as the numerical factors in the bounds are worse than ours. The case of a bounded regret is also not covered, while it constitutes our main contribution; see Section 1.3 below.

The main reason behind these suboptimal bounds is related to a loss of information, due to discarding the pure action A_t picked, which is known, and relating the reward $Y_t \sim \mu_{A_t}$ to $\underline{p}_t$ and not to A_t. The considered setting can thus be described as a stochastic linear bandit setting with augmented feedback. See Appendix C for more details on these statements, including an intuition on why sharper regret bounds are achieved with the additional information of which arm A_t was picked and a literature review on bandit models with augmented feedback (and thus, improved regret bounds), discussing contributions like the ones by Caron et al. (2012) and Degenne et al. (2018).

1. In this article, we define finite polytopes as convex hulls of a finite set of points (definitions of polytopes vary by articles and there is no universal terminology).

1.3. Summary of our contributions and outline of the article

Section 2 introduces our main algorithm, a diversity-preserving variant of UCB, which is computationally efficient. Theorem 1 provides several regret bounds, the most interesting ones guaranteeing a bounded regret in the natural cases when some probability mass is put on all arms either just by the optimal mixed actions (non-explicit bound) or by all actions (closed-form bound); this corresponds to the cases when diversity is indeed desirable. Section 3 discusses lower bounds: Theorem 2, which relies on the approach introduced by Graves and Lai (1997), indicates that $\ln T$ rates are unavoidable in the case when some arms receive zero probability mass by optimal actions, if in addition the means of the distributions in $\mathcal{D}$ are not bounded from above. Section 4 illustrates the dual behavior of either a bounded regret or a $\ln T$ regret for our variant of UCB. Most of our claims are proved in the main body of this article. However, an appendix collects an extended literature review and technical considerations used in the proofs of Sections 2 and 3. It also briefly discusses distribution-free regret bounds (which are less challenging in this diversity-preserving setting), see Appendix A.

2. Diversity-preserving UCB: distribution-dependent regret upper bounds

To state our main algorithm, we first introduce estimations of the means μ_a of the arms $a \in [K]$. We define

$$N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s = a\}} \qquad \text{and} \qquad \widehat{\mu}_a(t) = \begin{cases} 1 & \text{if } N_a(t) = 0 \\ \dfrac{1}{N_a(t)} \displaystyle\sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s = a\}} & \text{if } N_a(t) \geqslant 1 \end{cases}$$

Note that in the diversity-preserving setting, we cannot ensure that arm a be picked even once. Therefore, contrary to the vanilla bandit setting, it is important to handle the case when $N_a(t) = 0$. We thus set a default value of 1 (the maximal average reward) for $\widehat{\mu}_a(t)$ in this case. For the same reason, we put a maximum in the denominator of the upper confidence bounds $U_a(t)$, see Algorithm 1 below.

We assume that $\mathcal{P}$ is a finite polytope (see Footnote 1): it is the convex hull of a finite set of points $\mathrm{Ext}(\mathcal{P})$. The natural extension of the UCB algorithm to our setting is stated next. Note that the maximum of the linear functional $\underline{p} \in \mathcal{P} \mapsto \langle \underline{p}, \underline{U}(t-1) \rangle$ is reached for some $\underline{p}$ in $\mathrm{Ext}(\mathcal{P})$. The requirement that $\underline{p}_t$ be chosen among $\mathrm{Ext}(\mathcal{P})$ only is made for technical reasons.

Algorithm 1 Diversity-preserving UCB for rewards in $[0,1]$ and when $\mathcal{P}$ is a finite polytope

1: **Initialization**: $\underline{U}(0) = (1, \dots, 1)$
2: **for** rounds $t = 1, \dots,$ **do**
3: Select (ties broken arbitrarily) and play $\underline{p}_t \in \underset{\underline{p} \in \mathrm{Ext}(\mathcal{P})}{\mathrm{argmax}} \langle \underline{p}, \underline{U}(t-1) \rangle$
4: Play the pure action $A_t \sim \underline{p}_t$
5: Get and observe the reward $Y_t \sim \nu_{A_t}$
6: Compute the upper confidence bound vector $\underline{U}(t) = \big(U_1(t), \dots, U_K(t)\big)$ according to

$$\forall a \in [K], \qquad U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2 \ln t}{\max\big\{N_a(t), 1\big\}}}$$

7: **end for**

The analysis relies on the suboptimality gaps. To define them, we first define the optimal expected payoff and the set of optimal mixed actions as

$$M(\underline{\mu}, \mathcal{P}) = \max_{\underline{p} \in \mathcal{P}} \langle \underline{p}, \underline{\mu} \rangle \qquad \text{and} \qquad \mathrm{Opt}(\underline{\nu}) = \operatorname*{argmax}_{\underline{p} \in \mathcal{P}} \langle \underline{p}, \underline{\mu} \rangle \,. \tag{1}$$

The suboptimality gap of a given mixed action $p \in \mathcal{P}$ and the suboptimality gap of the set $\mathrm{Ext}(\mathcal{P})$ are in turn defined as

$$\Delta(\underline{p}) = M(\underline{\mu}, \mathcal{P}) - \langle \underline{p}, \underline{\mu} \rangle \qquad \text{and} \qquad \Delta = \min\big\{ \Delta(\underline{p}) : \underline{p} \in \mathrm{Ext}(\mathcal{P}),\ \Delta(\underline{p}) > 0 \big\} \,. \tag{2}$$

(We assume that at least one $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ is such that $\Delta(\underline{p}) > 0$, otherwise, all strategies have a null regret.) In the analysis, it will be handy to consider, for $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ and $t \geqslant 1$,

$$N_{\underline{p}}(t) = \sum_{s=1}^{t} \mathbb{1}_{\{\underline{p}_s = \underline{p}\}} \,, \qquad \text{so that (by the tower rule)} \qquad R_T(\underline{\nu}) = \sum_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} \Delta(\underline{p}) \, \mathbb{E}\big[N_{\underline{p}}(T)\big] \,. \tag{3}$$

In this setting, that decomposition of the regret in terms of gaps of mixed actions $\underline{p}$ is more useful than the classical decomposition in terms of gaps for each pure action $a \in [K]$.

To state our main bound, we define some minimal values of probabilities of playing each arm a, as far as optimal mixed actions and all mixed actions are concerned, respectively:

$$p^\star_{\min}(\underline{\nu}) = \min_{\underline{p} \in \mathrm{Opt}(\underline{\nu})} \min_{a \in [K]} p_a \qquad \text{and} \qquad \ell = \min_{\underline{p} \in \mathcal{P}} \min_{a \in [K]} p_a \,.$$

The fact that $p^\star_{\min}(\underline{\nu}) > 0$ (respectively, $\ell > 0$) corresponds to the case when $\mathrm{Opt}(\underline{\nu})$ (respectively, $\mathcal{P}$) is in the relative interior of the simplex. The assumption $\ell > 0$ of Case 3 in the theorem below is more stringent than the assumption $p^\star_{\min}(\underline{\nu}) > 0$ of Case 2 (and Case 1 comes with no assumption).

Theorem 1 *Assume that the diversity-preserving set $\mathcal{P}$ is a finite polytope and the bandit model is $\mathcal{D}_{[0,1]}$, the set of all distributions over $[0,1]$. Then, the regret of diversity-preserving UCB (Algorithm 1) is bounded as follows, for all bandit problems $\underline{\nu}$:*

1. *In all cases,* $\displaystyle R_T(\underline{\nu}) \leqslant \frac{24K\big(\ln(1+T)\big)^2 + K + 2}{\Delta} \,.$
2. *If $\mathcal{P}$ and $\underline{\nu}$ are such that $p^\star_{\min}(\underline{\nu}) > 0$, then the regret is bounded,* $\displaystyle \lim_{T \to \infty} R_T(\underline{\nu}) < +\infty \,.$
3. *If $\mathcal{P}$ is such that $\ell > 0$, then the regret is bounded by the closed-form bound*

$$R_T(\underline{\nu}) \leqslant \frac{24K}{\Delta} \ln\left(1 + \frac{32}{\Delta^2 \ell} \ln\left(\frac{16}{\Delta^2 \ell}\right)\right) + \frac{7K+3}{\min\{\Delta, \ell^2\}} \,.$$

The theorem is proved in the rest of this section. For now, we issue two series of comments.
Comment 1: The assumption of Case 2, which reads $p^\star_a > 0$ for all $a \in [K]$ and all optimal mixed actions $\underline{p}^\star$, is a natural assumption, which models the fact that preserving diversity is desirable (all arms are somewhat useful). We show that bounded regret is possible in this case.
Comment 2: Of course, running a classical UCB on the mixed actions in $\mathrm{Ext}(\mathcal{P})$ would guarantee a regret bound of order $\big(\sum_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} 1/\Delta(\underline{p})\big) \ln T$. This would even improve asymptotically over the $(\ln T)^2$ rate of Case 1. However it would come at a price of huge impractical constants when $\mathcal{P}$ has many vertices, and would not cover the case of bounded regret. Importantly, our algorithm is also more elegant and computationally slightly more efficient.

2.1. Common part of the proofs

We consider the events

$$\mathcal{E}(t) = \left\{ \text{For all } a \in [K], \quad \left|\mu_a - \widehat{\mu}_a(t)\right| \leqslant \sqrt{\frac{2\ln t}{\max\{N_a(t), 1\}}} \right\} \tag{4}$$

$$\text{and} \qquad \mathcal{E}'(t) = \left\{ \text{For all } a \in [K], \quad \sqrt{\frac{2\ln t}{\max\{N_a(t), 1\}}} < \frac{\Delta}{2} \right\}.$$

When both $\mathcal{E}(t)$ and $\mathcal{E}'(t)$ hold, for any suboptimal $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ and any optimal mixed action $\underline{p}^\star$, we have the following chain of inequalities, where we use $\Delta \leqslant \Delta(\underline{p})$:

$$\langle \underline{U}(t), \underline{p} \rangle = \langle \widehat{\mu}(t), \underline{p} \rangle + \sum_{a=1}^{K} p_a \sqrt{\frac{2\ln t}{\max\{N_a(t), 1\}}} \leqslant \langle \underline{\mu}, \underline{p} \rangle + 2\sum_{a=1}^{K} p_a \sqrt{\frac{2\ln t}{\max\{N_a(t), 1\}}}$$

$$< \langle \underline{\mu}, \underline{p} \rangle + \Delta \leqslant \langle \underline{\mu}, \underline{p} \rangle + \Delta(\underline{p}) = \langle \underline{\mu}, \underline{p}^\star \rangle \leqslant \langle \widehat{\underline{\mu}}(t), \underline{p}^\star \rangle + \sum_{a=1}^{K} p_a^\star \sqrt{\frac{2\ln t}{\max\{N_a(t), 1\}}} = \langle \underline{U}(t), \underline{p}^\star \rangle .$$

In that case, by construction of our algorithm as an index policy, no suboptimal mixed action is picked; put differently, and writing $\overline{B}$ for the complement of any event B,

$$\left\{ p_{t+1} \notin \mathrm{Opt}(\underline{\nu}) \right\} \subseteq \overline{\mathcal{E}(t)} \cup \overline{\mathcal{E}'(t)} .$$

Since the (non-expected) instantaneous regrets $r_t = \langle \underline{p}^\star - \underline{p}_t \,, \underline{\mu} \rangle$ are always smaller than 1 given that the considered bandit problem lies in $\mathcal{D}_{[0,1]}$, and for a time $t_0 \geqslant 1$ to be defined by the analysis,

$$\begin{aligned} R_T(\underline{\nu}) &= R_{t_0}(\underline{\nu}) + \sum_{t=t_0+1}^{T} \mathbb{E}[r_t] \leqslant R_{t_0}(\underline{\nu}) + \sum_{t=t_0+1}^{T} \mathbb{P}\left\{ p_t \notin \mathrm{Opt}(\underline{\nu}) \right\} \\ &\leqslant R_{t_0}(\underline{\nu}) + \sum_{t=t_0}^{T-1} \underbrace{\mathbb{P}\Big(\overline{\mathcal{E}(t)}\Big)}_{\leqslant 2Kt^{-3}} + \sum_{t=t_0}^{T-1} \mathbb{P}\Big(\overline{\mathcal{E}'(t)}\Big) \leqslant R_{t_0}(\underline{\nu}) + 3K + \sum_{t=t_0}^{T-1} \mathbb{P}\Big(\overline{\mathcal{E}'(t)}\Big) , \end{aligned} \tag{5}$$

where, for the final inequality, we used Lemma 4 of Appendix D (which is a direct application of Hoeffding's inequality together with a union bound) to get $\mathbb{P}\big(\overline{\mathcal{E}(t)}\big) \leqslant 2Kt^{-3}$ for $t \geqslant 2$ (true also for $t = 1$), together with the fact that the series of the $1/t^3$ sums up over $t \geqslant 1$ to $\zeta(3) < 1.21$.

2.2. Proof of Case 3 of Theorem 1

Let $t_0 = \max\big\{ t \in \{2, 3, 4, \ldots\} : (8\ln t)/\Delta^2 > t\ell/2 \big\}$, which is well defined since $(8\ln 2)/\Delta^2 > 1$ and $(\ln t)/t \to 0$ as $t \to +\infty$. We have, for $t \geqslant t_0 + 1$,

$$\overline{\mathcal{E}'(t)} = \left\{ \exists a \in [K] : \; N_a(t) \leqslant \frac{8\ln t}{\Delta^2} \right\} \subseteq \bigcup_{a \in [K]} \left\{ N_a(t) \leqslant \frac{t\ell}{2} \right\} \subseteq \bigcup_{a \in [K]} \left\{ N_a(t) - \sum_{s=1}^{t} p_{s,a} \leqslant -\frac{t\ell}{2} \right\} ,$$

where we used, for the last inclusion, the fact that $p_{s,a} \geqslant \ell$ by definition of ℓ. By the Hoeffding-Azuma inequality (see, e.g., Lemma A.7 in Cesa-Bianchi and Lugosi 2006), for all $t \geqslant 1$ and all $\varepsilon > 0$,

$$\mathbb{P}\left\{N_a(t) - \sum_{s=1}^{t} p_{s,a} \leqslant -\varepsilon\right\} \leqslant \exp\bigl(-2\varepsilon^2/t\bigr)\,.$$

Therefore,

$$\sum_{t=t_0}^{+\infty} \mathbb{P}\Bigl(\overline{\mathcal{E}'(t)}\Bigr) \leqslant 1 + \sum_{t=t_0+1}^{+\infty} \sum_{a\in[K]} \exp\left(-\frac{2}{t}\left(\frac{t\ell}{2}\right)^2\right) \leqslant 1 + K \sum_{t=t_0+1}^{+\infty} \exp\left(-\frac{t\ell^2}{2}\right) \leqslant 1 + \frac{K}{1-\mathrm{e}^{-\ell^2/2}}\,.$$

We substitute the above bound into (5) together with $R_{t_0} \leqslant 24K\bigl(\ln(1+t_0)\bigr)^2/\Delta + (2+K)/\Delta$, which follows from Case 1 of Theorem 1 (later proved in Appendix D.3); we get

$$R_T(\underline{\nu}) \leqslant \frac{24K\bigl(\ln(1+t_0)\bigr)^2}{\Delta} + 3K + \frac{2+K}{\Delta} + 1 + \frac{K}{1-\mathrm{e}^{-\ell^2/2}}\,.$$

The proof of Case 3 is concluded by proving (see Appendix D.1)

$$t_0 \leqslant \frac{32}{\Delta^2\ell} \ln\left(\frac{16}{\Delta^2\ell}\right) \tag{6}$$

and by performing crude boundings to improve readability, like $1/(1-\mathrm{e}^{-u}) \leqslant 1 + 1/u$ for $u > 0$.

2.3. Proof of Case 2 of Theorem 1

Based on (5) with $t_0 = 1$, we only need to prove that the sum of the $\mathbb{P}\Bigl(\overline{\mathcal{E}'(t)}\Bigr)$ over $t \geqslant 1$ is finite. To do so, we denote, for $t \geqslant 1$,

$$N_\star(t) = \sum_{\underline{p}\in\mathrm{Opt}(\underline{\nu})} N_{\underline{p}}(t)$$

the number of times up to round t an optimal mixed action was pulled. We decompose events based on whether $N_\star(t)$ is larger or smaller than $t/2$: by several applications of the union bound,

$$\begin{aligned}\mathbb{P}\Bigl(\overline{\mathcal{E}'(t)}\Bigr) &\leqslant \mathbb{P}\left(\left\{N_\star(t) \geqslant \frac{t}{2}\right\} \cap \overline{\mathcal{E}'(t)}\right) + \mathbb{P}\left\{N_\star(t) \leqslant \frac{t}{2}\right\} \\ &= \sum_{a=1}^{K} \mathbb{P}\left(\left\{N_\star(t) \geqslant \frac{t}{2}\right\} \cap \left\{N_a(t) \leqslant \frac{8\ln t}{\Delta^2}\right\}\right) + \mathbb{P}\left\{N_\star(t) \leqslant \frac{t}{2}\right\}\,. \end{aligned} \tag{7}$$

We prove in Appendix D.2 that optimal mixed actions are pulled an overwhelming fraction of the time, typically more than half of the time:

$$\sum_{t=1}^{+\infty} \mathbb{P}\left\{N_\star(t) \leqslant \frac{t}{2}\right\} < +\infty\,. \tag{8}$$

As for the other sum, we fix $a \in [K]$ and first note that by the Hoeffding-Azuma inequality, for each $t \geqslant 1$ and all $\delta_t \in (0,1]$,

$$\mathbb{P}\left\{N_a(t) - \sum_{s=1}^{t} p_{s,a} \leqslant -\sqrt{\frac{t}{2}\ln\frac{1}{\delta_t}}\right\} \leqslant \delta_t\,.$$

Note that $p_{s,a} \geqslant \mathbb{1}_{\{\underline{p}_s \in \mathrm{Opt}(\underline{\nu})\}}\, p^\star_{\min}(\underline{\nu})$ by definition of $p^\star_{\min}(\underline{\nu})$. Therefore, choosing $\delta_t = 1/t^2$,

$$\mathbb{P}\big\{N_a(t) - p^\star_{\min}(\underline{\nu})\, N_\star(t) \leqslant -\sqrt{t \ln t}\big\} \leqslant 1/t^2\,.$$

Now, as $-\sqrt{t\ln t} \geqslant (8\ln t)/\Delta^2 - t p^\star_{\min}(\underline{\nu})/2$ for all t larger than some t_0', we finally get, for $t \geqslant t_0'$,

$$\mathbb{P}\left(\left\{N_\star(t) \geqslant \frac{t}{2}\right\} \cap \left\{N_a(t) \leqslant \frac{8\ln t}{\Delta^2}\right\}\right) \leqslant \mathbb{P}\left\{N_a(t) - p^\star_{\min}(\underline{\nu})\, N_\star(t) \leqslant \frac{8\ln t}{\Delta^2} - \frac{t p^\star_{\min}(\underline{\nu})}{2}\right\} \leqslant 1/t^2\,.$$

This shows that the first sum in (7) is also bounded, which concludes the proof of this case.

2.4. Proof of Case 1 of Theorem 1

It follows closely the standard proof scheme for linear bandits, see Lattimore and Szepesvári (2020, Chapter 19) and earlier references therein; thanks to the knowledge of the pure actions A_t played we are able to simplify the proof and get a more readable bound, see Appendix D.3.

3. Distribution-dependent lower bounds for finite polytopes $\mathcal{P}$

This section considers general models $\mathcal{D}$ for the distributions of the considered bandit problems $\nu = (\nu_1, \dots, \nu_K)$, with finite first moments, and satisfying a mild assumption (see below). Our aim is to discuss the optimality of the regret upper bounds exhibited in the previous section. To do so, we rely on and adapt the approach introduced by Graves and Lai (1997).

We first describe the assumptions that we make in this section. We only consider diversity-preserving sets $\mathcal{P}$ given by *finite polytopes*, i.e., sets which are the convex hull of a finite set of points denoted by $\mathrm{Ext}(\mathcal{P})$. Recall from (1) that $\mathrm{Opt}(\underline{\nu})$ refers to the set of optimal mixed actions of a bandit problem $\underline{\nu}$. We will assume that there is a *unique optimal mixed action*: $\mathrm{Opt}(\underline{\nu}) = \big\{\underline{p}^\star(\underline{\nu})\big\}$; this mixed action then necessarily belongs to $\mathrm{Ext}(\mathcal{P})$. This assumption is common in bandit analyses (it is also made, e.g., in Lattimore and Szepesvári, 2017, Combes et al., 2017), and it is arguably harmless as generic problems will typically have a unique optimal mixed action. Another, harmless, assumption is that $\mathcal{P}$, or equivalently, $\mathrm{Ext}(\mathcal{P})$, *puts some probability mass on each arm* $a \in [K]$, that is: for each $a \in [K]$, there exists a mixed action $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ with $p_a > 0$. If this is not the case, then no mixed action in $\mathcal{P}$ puts a positive mass on a and a will never be played, thus should be discarded.

Finally, we make the following technical assumption on $\mathcal{D}$, where the set of confusing alternative problems $\mathrm{ALT}(\underline{\nu})$ is defined in (9) below.

Assumption 3.1 *For all $\underline{\nu}$ in $\mathcal{D}$, the set $\mathrm{ALT}(\underline{\nu})$ defined in (9) is either empty or contains some $\underline{\nu}''$ such that $\mathrm{KL}(\nu_a, \nu_a'') < +\infty$ for all $a \in [K]$.*

This technical assumption is immediately satisfied in many common situations: when $\mathcal{D}$ is a convex subset of the set of all probability distributions on the real line (see Appendix E.3 for a proof) or when $\mathcal{D}$ is such that $\mathrm{KL}(\nu_0, \nu_0'') < +\infty$ for all distributions ν_0 and ν_0'' in $\mathcal{D}$, as is the case, for instance, for canonical exponential families indexed by an open interval (this follows from the closed-form expression of KL in that case, see, e.g., Garivier and Cappé, 2011, Lemma 6).

Given the regret upper bounds exhibited in Section 2, it is natural to restrict our attention to *uniformly fast convergent [UFC]* strategies over $\mathcal{D}$ given the diversity-preserving set $\mathcal{P}$: by definition, such strategies ensure that for any bandit problem $\underline{\nu}'$ in $\mathcal{D}$, their diversity-preserving regret satisfies $R_T(\underline{\nu}') = o(T^\alpha)$ for all $\alpha > 0$.

In this section (and its corresponding Appendix E), we index the regret by the underlying bandit problem $\underline{\nu}$ for the sake of clarity and write $R_T(\underline{\nu})$. Whenever needed, we also index expectations $\mathbb{E}$ by the underlying bandit problem $\underline{\nu}$, by writing $\mathbb{E}_{\underline{\nu}}$.

Theorem 2 *Assume that the diversity-preserving set $\mathcal{P}$ is a finite polytope, generated by the finite set $\mathrm{Ext}(\mathcal{P})$, and that it puts some probability mass on each arm $a \in [K]$. For all models $\mathcal{D}$ with finite first moments and satisfying Assumption 3.1, for all strategies that are uniformly fast convergent [UFC] over $\mathcal{D}$ given $\mathcal{P}$, for all bandit problems $\underline{\nu} = (\nu_1, \dots, \nu_K)$ in $\mathcal{D}$ with a unique optimal mixed action $\underline{p}^\star(\underline{\nu})$,*

$$\liminf_{T\to\infty} \frac{R_T(\underline{\nu})}{\ln T} \geqslant c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big)\,,$$

where $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big) \in [0, +\infty)$ is defined below in Definition 3 as a constrained infimum.

If $\underline{p}^\star(\underline{\nu})$ is such that $p^\star_a(\underline{\nu}) > 0$ for all $a \in [K]$, then $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big) = 0$. The converse implication, that is, the fact that $p^\star_a(\underline{\nu}) = 0$ for some $a \in [K]$ entails $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big) > 0$, also holds when the means of the distributions in $\mathcal{D}$ are not bounded from above.

The most interesting part of this theorem is its second part. We know from Case 2 of Theorem 1 that bounded regret is achievable when $\mathcal{D} = \mathcal{D}_{[0,1]}$ and $p^\star_a > 0$ for all $a \in [K]$, which is consistent with $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big) = 0$. Theorem 2 indicates that the $\ln T$ rates are unavoidable in the case when $p^\star_a = 0$ for some $a \in [K]$ and when the means of the distributions in $\mathcal{D}$ are not bounded from above, which is not the case for the model $\mathcal{D}_{[0,1]}$. Theorems 1 and 2 could cover a common case given by a model $\mathcal{D}_{\sigma^2-\mathrm{SG}}$ composed sub-Gaussian distributions with known variance factor σ^2 with no upper bound on the means; we indeed feel that Theorem 1 could be extended to also cover that case. If so, we would have a true dual behavior for the regret: either bounded or growing as $\ln T$.

3.1. Proof of Theorem 2

We provide a rather detailed sketch of proof; two omitted arguments may be found in Appendix E.

Reduction argument. The first step of the proof consists in noting that any strategy picking mixed distributions in the finite polytope $\mathcal{P}$ can be converted into a (randomized) strategy picking mixed distributions in the finite set $\mathrm{Ext}(\mathcal{P})$ only and providing the same expected cumulative reward, thus suffering the same expected regret R_T. This is achieved by an extra randomization step and crucially relies on the fact that any strategy ultimately needs to play a pure action A_t at each step (this reduction would not work in linear bandits). Details are to be found in Appendix E.1. We therefore only need to prove the lower bound for (possibly randomized) strategies playing in the finite set $\mathrm{Ext}(\mathcal{P})$.

First part of the theorem: à la Graves and Lai (1997). Some of the notation used below was defined earlier in the article, e.g., $\mathrm{Opt}(\underline{\nu}')$ and $\Delta(\underline{p})$ were defined in (1) and (2), respectively.

We introduce the set of confusing alternative problems associated with the bandit problem $\underline{\nu}$, denoted by $\mathrm{ALT}(\underline{\nu})$. Problems in $\mathrm{ALT}(\underline{\nu})$ are the ones in which $p^\star(\underline{\nu})$ is suboptimal, but that the

player cannot discriminate from $\underline{\nu}$ by only playing $p^\star(\underline{\nu})$. Formally, for each arm a, either $p_a^\star(\underline{\nu}) = 0$ and selecting the optimal probability $\underline{p}^\star(\underline{\nu})$ never results in picking arm a, or $\nu_a = \nu'_a$ and observing a reward associated with a does not provide discriminative information:

$$\textsc{Alt}(\underline{\nu}) = \left\{ \underline{\nu}' \text{ in } \mathcal{D} \;\middle|\; p^\star(\underline{\nu}) \notin \mathrm{Opt}(\underline{\nu}') \;\text{ and }\; \forall\, 1 \leqslant a \leqslant K, \quad p_a^\star(\underline{\nu}) = 0 \;\text{ or }\; \nu_a = \nu'_a \right\}. \tag{9}$$

Thanks to Theorem 1, we know that the correct scaling of the suboptimal pulls is at most logarithmic; we therefore define the normalized allocations, for all $\underline{p} \in \mathrm{Ext}(\mathcal{P})$,

$$n_T(\underline{p}) = \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\underline{p}}(T)\big]}{\ln T}, \qquad \text{so that} \quad \frac{R_T}{\ln T} = \sum_{\underline{p}\in\mathcal{P}} \Delta(\underline{p}) \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\underline{p}}(T)\big]}{\ln T} = \sum_{\underline{p}\in\mathcal{P}} \Delta(\underline{p})\, n_T(\underline{p}). \tag{10}$$

A UFC algorithm facing the problem $\underline{\nu}$ will eventually focus on the unique optimal mixed action $\underline{p}^\star(\underline{\nu})$. Because of this, most of its observations will correspond to pure actions $a \in [K]$ such that $p_a^\star(\underline{\nu}) > 0$, which provide no information that is useful to distinguish $\underline{\nu}$ from problems $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$. A measure of this useful information is the Kullback-Leibler divergence between the distributions of arms $A_1, \dots, A_T$ picked and the rewards $Y_1, \dots, Y_T$ obtained in the first T rounds, $\mathbb{P}_{\underline{\nu},T}$ and $\mathbb{P}_{\underline{\nu}',T}$, when the underlying problems are $\underline{\nu}$ and $\underline{\nu}'$, respectively. It may be computed thanks to a chain rule, see Equation (9) in Garivier et al. (2019) for the first equality below, followed by an application of the tower rule for the second equality:

$$\mathcal{I}_T = \mathrm{KL}\left(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\right) = \sum_{t=1}^{T} \mathbb{E}_{\underline{\nu}}\big[\mathrm{KL}(\nu_{A_t}, \nu'_{A_t})\big] = \sum_{t=1}^{T} \mathbb{E}_{\underline{\nu}}\left[\sum_{a=1}^{K} p_{t,a}\, \mathrm{KL}(\nu_a, \nu'_a)\right]. \tag{11}$$

This quantity can be factored as a sum over the mixed actions $\underline{p} \in \mathrm{Ext}(\mathcal{P})$:

$$\mathcal{I}_T = \sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} \mathbb{E}_{\underline{\nu}}\big[N_{\underline{p}}(T)\big] \sum_{a\in[K]} p_a\, \mathrm{KL}(\nu_a, \nu'_a) = (\ln T) \sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} n_T(\underline{p}) \sum_{\substack{a\in[K] \\ p_a^\star(\underline{\nu})=0}} p_a\, \mathrm{KL}(\nu_a, \nu'_a), \tag{12}$$

where the final equality holds as by definition, problems $\nu' \in \textsc{Alt}(\underline{\nu})$ are such that $\nu'_a = \nu_a$ when $p_a^\star(\underline{\nu}) > 0$. Asymptotically, the algorithm must maintain this amount of information above $\ln T$ in order to satisfy the UFC assumption, see details in Appendix E. This puts a constraint on the limit of $n_T(\underline{p})$ for all $\underline{p}$, which may be read in Equation (13) below. Given the rewriting (10) of the regret we then get the following definition for the quantity $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big)$ contemplated in Theorem 2.

Definition 3 *The constrained infimum $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big)$ in Theorem 2 is defined as:*

$$\inf_{n\in\mathbb{R}_+^{\mathrm{Ext}(\mathcal{P})}} \sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} \Delta(\underline{p})\, n(\underline{p}) \quad \textit{under the constraint that}$$

$$\forall\, \underline{\nu}' \in \textsc{Alt}(\underline{\nu}), \quad \sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P}) \\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n(\underline{p}) \sum_{\substack{a\in[K] \\ p_a^\star(\underline{\nu})=0}} p_a\, \mathrm{KL}(\nu_a, \nu'_a) \geqslant 1. \tag{13}$$

We conveyed some intuition on the lower bound indicated by the first part of Theorem 2 and could state the definition of $c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big)$. The rest of the proof, which follows standard techniques introduced by Graves and Lai (1997), may be found in Appendix E.2. We now turn to the second part of Theorem 2.

Second part of the theorem: checking whether $\textsc{Alt}(\underline{\nu})$ is empty or not. We rely on the following equivalence: in the setting and under the conditions of Theorem 2,

$$c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)=0 \qquad \Longleftrightarrow \qquad \textsc{Alt}(\underline{\nu})=\emptyset\,. \tag{14}$$

Proof of (14). Indeed, if $\textsc{Alt}(\underline{\nu})=\emptyset$, then the linear program (13) is unconstrained and yields $c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)=0$. If $\textsc{Alt}(\underline{\nu})$ is non-empty, by Assumption 3.1, there exists at least one $\underline{\nu}''\in\textsc{Alt}(\underline{\nu})$ such that $\mathrm{KL}(\nu_a,\nu_a'')<+\infty$ for all $a\in[K]$, which we fix. For a vector $n\in\mathbb{R}_+^{\mathrm{Ext}(\mathcal{P})}$ to satisfy the constraint (13), it is necessary that

$$\sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a\,\mathrm{KL}(\nu_a,\nu_a'')\geqslant 1\,. \tag{15}$$

Since

$$\sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a\,\mathrm{KL}(\nu_a,\nu_a'')\leqslant C_{\underline{\nu},\underline{\nu}''}\sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}}\Delta(\underline{p})n(\underline{p})\,,$$

$$\text{where}\qquad C_{\underline{\nu},\underline{\nu}''}=\max_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}}\frac{1}{\Delta(\underline{p})}\sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a\,\mathrm{KL}(\nu_a,\nu_a'')\leqslant\frac{1}{\Delta}\max_{a\in[K]}\mathrm{KL}(\nu_a,\nu_a'')<+\infty\,,$$

the constraint (15) entails that

$$\sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})}\Delta(\underline{p})\,n(\underline{p})\geqslant\frac{1}{C_{\underline{\nu},\underline{\nu}''}}>0\,,$$

proving that $c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)\geqslant 1/C_{\underline{\nu},\underline{\nu}''}>0$.

Exploitation of (14). If $p_a^\star>0$, then the only $\underline{\nu}'$ such that $p_a^\star=0$ or $\nu_a=\nu_a'$ for all $a\in[K]$ is $\underline{\nu}$ itself; that is, $\textsc{Alt}(\underline{\nu})=\emptyset$ and therefore, by (14), we have $c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)=0$ as stated in the second part of Theorem 2.

We now assume that the means of the distributions in $\mathcal{D}$ are not bounded from above and show that $p_a^\star(\underline{\nu})=0$ for some $a\in[K]$ entails that $\textsc{Alt}(\underline{\nu})$ is non-empty, thus $c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)>0$ by (14). We fix such an a. By assumption, $\mathcal{P}$ thus $\mathrm{Ext}(\mathcal{P})$ put some probability mass on this arm a: there exists $\underline{p}\in\mathrm{Ext}(\mathcal{P})$ with $p_a>0$. Since $\underline{p}^\star(\nu)$ is the unique optimal arm of ν, the gap $\Delta(\underline{p})$ is positive. Now, by the assumption of unbounded means in $\mathcal{D}$, there exists a distribution $\nu_a'\in\mathcal{D}$ with expectation $\mu_a'>\mu_a+\Delta(\underline{p})/p_a$. We denote by $\underline{\nu}'$ the bandit problem such that $\nu_k'=\nu_k$ for all $k\neq a$, and whose a–th distribution is ν_a'. The mixed action $\underline{p}^\star(\underline{\nu})$ is suboptimal for $\underline{\nu}'$: we have indeed, by construction of all quantities, by definition of $\Delta(\underline{p})$, and since $p_a^\star(\underline{\nu})=0$ while $\underline{\nu}$ and $\underline{\nu}'$ only differ at a,

$$\big\langle\underline{p},\,\underline{\mu}'\big\rangle=\big\langle\underline{p},\,\underline{\mu}\big\rangle+(\mu_a'-\mu_a)\,p_a>\big\langle\underline{p},\,\underline{\mu}\big\rangle+\frac{\Delta(\underline{p})}{p_a}\,p_a=\big\langle\underline{p},\,\underline{\mu}\big\rangle+\Delta(\underline{p})=\big\langle\underline{p}^\star(\underline{\nu}),\,\underline{\mu}\big\rangle=\big\langle\underline{p}^\star(\underline{\nu}),\,\underline{\mu}'\big\rangle\,.$$

That is, $p^\star(\underline{\nu})\notin\mathrm{Opt}(\underline{\nu}')$. We thus proved that $\underline{\nu}'\in\textsc{Alt}(\underline{\nu})$, so that $\textsc{Alt}(\underline{\nu})$ is non-empty.

4. Some experiments on synthetic data

In this section, we perform some experiments that illustrate the dual behavior of the regret: either bounded or growing at a $\ln T$ rate. More precisely, we consider bandit problems $\underline{\nu}$ with a unique optimal mixed action $\underline{p}^\star(\underline{\nu})$ and illustrate that in the Bernoulli model considered, either a $\ln T$ rate for regret is suffered when $p_a^\star(\underline{\nu}) = 0$ for some arm $a \in [K]$, while a bounded regret is achieved when $p_a^\star(\underline{\nu}) > 0$ for all arms $a \in [K]$.

Setting considered. We consider $K = 3$ arms and the diversity-preserving set

$$\mathcal{P}_\ell = \big\{(p_1, p_2, p_3) \in \mathcal{S} \ : \quad p_1 \geqslant \ell \quad \text{and} \quad p_2 \geqslant \ell\big\}\,,$$

where $\ell \in (0, 1/2)$ is a parameter. The set $\mathcal{P}_\ell$ is a finite polytope generated by

$$\underline{p}^{(1)} = \big(\ell,\, 1-\ell,\, 0\big), \qquad \underline{p}^{(2)} = \big(1-\ell,\, \ell,\, 0\big), \qquad \text{and} \qquad \underline{p}^{(1,2)} = \big(\ell,\, \ell,\, 1-2\ell\big)\,.$$

The model $\mathcal{D}$ is given by Bernoulli distributions. We will consider bandits problems $\underline{\nu}_\alpha$ in $\mathcal{D}$, each of them indexed by $\alpha \in (-1/6, 1/6)$. For $\alpha < 0$, the unique optimal mixed action will be $\underline{p}^{(1,2)}$, which satisfies $p_a^{(1,2)} > 0$ for all $a \in [3]$, and bounded regret will be achieved. For $\alpha > 0$, the unique optimal mixed action will be $\underline{p}^{(2)}$, which satisfies $p_3^{(2)} = 0$, and a $\ln T$ regret will be illustrated. More precisely,

$$\underline{\nu}_\alpha = \big(\mathrm{Ber}(1/2+\alpha),\, \mathrm{Ber}(1/3),\, \mathrm{Ber}(1/2-\alpha)\big)\,, \qquad \text{with} \qquad \underline{\mu}_\alpha = (1/2+\alpha,\, 1/3,\, 1/2-\alpha)\,.$$

The mixed action $\underline{p}^{(1)}$ is always dominated by $\underline{p}^{(2)}$ and $\underline{p}^{(1,2)}$: for all $\alpha \in (-1/6, 1/6)$,

$$\begin{aligned}
\big\langle \underline{p}^{(2)} - \underline{p}^{(1)},\, \underline{\mu}_\alpha \big\rangle &= (1-2\ell)\Big(\frac{1}{2}+\alpha\Big) + (2\ell-1)\frac{1}{3} + 0 = (1-2\ell)\Big(\frac{1}{6}+\alpha\Big) > 0\,,\\
\big\langle \underline{p}^{(1,2)} - \underline{p}^{(1)},\, \underline{\mu}_\alpha \big\rangle &= 0 + (2\ell-1)\frac{1}{3} + (1-2\ell)\Big(\frac{1}{2}-\alpha\Big) = (1-2\ell)\Big(\frac{1}{6}-\alpha\Big) > 0\,.
\end{aligned}$$

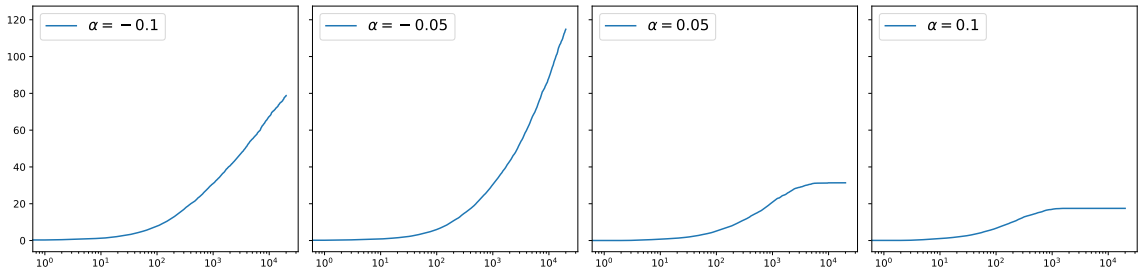
We now compare the mixed actions $\underline{p}^{(2)}$ and $\underline{p}^{(1,2)}$:

$$\big\langle \underline{p}^{(1,2)} - \underline{p}^{(2)},\, \underline{\mu}_\alpha \big\rangle = \big(2\ell-1\big)\Big(\frac{1}{2}+\alpha\Big) + 0 + \big(1-2\ell\big)\Big(\frac{1}{2}-\alpha\Big) = -2\alpha(1-2\ell)\,.$$

Numerical experiments. We set $\ell = 0.1$ and let α vary in $\{-0.1,\, -0.05,\, 0.05,\, 0.1\}$. We run the diversity-preserving UCB algorithm (Algorithm 1) on each of these four problems $\underline{\nu}_\alpha$, over $T = 20{,}000$ time steps, for $N = 75$ runs. The expected regret suffered by the algorithm is estimated by the empirical pseudo-regrets observed on the $N = 75$ runs:

$$\widehat{R}_T(\underline{\nu}_\alpha) = \frac{1}{N}\sum_{i=1}^{N} \widehat{R}_T(\underline{\nu}_\alpha, i)\,, \quad \text{where} \quad \widehat{R}_T(\underline{\nu}_\alpha, i) = \sum_{t=1}^{T} \big\langle \underline{p}^\star(\underline{\nu}_\alpha) - \underline{p}_t(\alpha, i),\, \underline{\mu}_\alpha \big\rangle\,,$$

and where we denoted by $\underline{p}_t(\alpha, i)$ the mixed action chosen at round t, during the i–th run, and for problem $\underline{\nu}_\alpha$. The figures below report the estimations obtained (solid lines); we also shaded areas corresponding to ± 2 standard errors of the estimates. As expected, the algorithm yields logarithmic regret when $\alpha < 0$ (the optimal mixed action is on the border of the simplex) and bounded regret when $\alpha > 0$ (the optimal mixed action lies in the interior of the simplex).



Acknowledgments

The work of Sébastien Gerchinovitz and Jean-Michel Loubes has benefitted from the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-PI3A-0004. Sébastien Gerchinovitz gratefully acknowledges the support of the DEEL project (https://www.deel.ai/).

References

Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS '11)*, pages 2312–2320, 2011.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems (NeurIPS '19)*, pages 9252–9262, 2019.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI '12)*, UAI'12, pages 142–151, 2012.

L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 160–169. Association for Computing Machinery, 2019.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. volume 124 of *Proceedings of Machine Learning Research*, pages 181–190. PMLR, 2020.

Yuan Shih Chow and Henry Teicher. *Probability Theory*. Springer, 1988.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AIStats'11)*, pages 208–214, 2011.

Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Reinforcement learning with fairness constraints for resource distribution in human-robot teams, 2019.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS'17)*, pages 1763–1771, 2017.

Rémy Degenne, Evrard Garcelon, and Vianney Perchet. Bandits with side observations: Bounded vs. logarithmic regret. arXiv 1807.03558, 2018.

Joseph L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory (COLT'11)*, pages 359–376, 2011.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems (NIPS'18)*, pages 2600–2609, 2018.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

Hédi Hadiji. *Sur quelques questions d'adaptation dans des problèmes de bandits stochastiques*. PhD thesis, 2020. URL https://www.imo.universite-paris-saclay.fr/~hadiji/files/manuscrit-26sept.pdf. The title can be translated into *Adaptive stochastic bandits*.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AIStats'20)*, volume 108, pages 3536–3545, 2020.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems (NIPS'16)*, pages 325–333, 2016.

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. arXiv 2006.08754, 2020.

Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems (NIPS'14)*, pages 550–558, 2014.

Tor Lattimore and Csaba Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AIStats'17)*, pages 728–737, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 2019.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 661–670, 2010.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. Calibrated fairness in bandits. In *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (Fat/ML 2017)*, 2017.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. arXiv 1907.10516, 2019.

Andrea Tirinzoni, Alessandro Lazaric, and Marcello Restelli. A novel confidence-based algorithm for structured bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AIStats'20)*, volume 108, 2020.

Outline of the appendix

Appendix A discusses distribution-free regret upper bounds.
Appendix B provides an extended literature review on the notions of diversity and fairness in bandits.
Appendix C is an extended comparison to the linear bandit setting used by Celis et al. (2019) and also provides a literature review on structured stochastic bandits.
Appendix D includes all remaining parts of the proofs of the regret upper bounds of Theorem 1.
Appendix E does so for the lower bound of Theorem 2.

Appendix A. Distribution-free regret upper bounds

We say that a strategy enjoys a distribution-free regret bound $B(\mathcal{D}, \mathcal{P}, T)$ on the model $\mathcal{D}$ given the diversity-preserving set $\mathcal{P}$ when it guarantees, for all $T \geqslant 1$,

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}} R_T(\underline{\nu}) \leqslant B(\mathcal{D}, \mathcal{P}, T)\,.$$

All claims that follow are extracted from Hadiji (2020, Chapter 5).

The diversity-preserving UCB (Algorithm 1) enjoys a distribution-free regret bound of order $\sqrt{KT \ln T}$, while a variant based on the MOSS strategy of Audibert and Bubeck (2010) achieves a $\sqrt{KT}$ bound.

A follow-the-regularizer-leader [FTRL] approach, similar to the one considered by Chen et al. (2020) and relying on a regularization function

$$H(\underline{p}) = \sum_{a \in [K]} p_a \ln p_a\,,$$

achieved a distribution-free regret bound of order

$$\sqrt{\operatorname{diam}_H(\mathcal{P})\, KT}\,, \qquad \text{where} \qquad \operatorname{diam}_H(\mathcal{P}) = \max\bigl\{H(\underline{p}) - H(\underline{q}) : \underline{p}, \underline{q} \in \mathcal{P}\bigr\}$$

is the diameter of $\mathcal{P}$ for the regularizer H. (It is in particular smaller than $\ln K$.) This FTRL bound actually also holds for adversarial bandits. Our results are incomparable with the ones by Chen et al. (2020): on the one hand, we consider more general diversity-preserving sets (they essentially consider the first and simplest example of Section 1.1, with minimal probabilities); on the other hand, they consider a contextual bandit setting.

A distribution-free regret lower bound of order

$$L_{\mathcal{P}}\sqrt{KT}\,, \qquad \text{where} \qquad L_{\mathcal{P}} = \left(\frac{1}{K}\sum_{i=1}^{K} \max_{p \in \mathcal{P}} p_i - \frac{1}{K}\right)^2,$$

may be proved, using standard techniques from Auer et al. (2002).

Statement of an open question. The optimal dependency of the distribution-free regret bounds on K and T is therefore $\sqrt{KT}$ for worst-case $\mathcal{P}$. However we were unable so far to identify the optimal dependency on the geometry of $\mathcal{P}$ in general.

Appendix B. Literature review on the notions of diversity and fairness in bandits

Diversity. Closest to our work is the recent article of Chen et al. (2020), that studies a particular case of the diversity-preserving setting, which essentially corresponds to our example called "Maintaining a budget" described in Section 1. While their framework is the same as ours, they only study distribution-free bounds and provide numerical experiments, making their focus orthogonal to our one. Note also that their algorithm, inspired by the mirror descent framework, actually applies to the rewards generated in an adversarial manner.

In the domain of stochastic bandits, we may cite the contributions of Patil et al. (2019) and Claure et al. (2019). They derive bandit algorithms that ensure that the proportion of times each action is selected is lower bounded, i.e., with our notation that $N_a(T)/T \geqslant \alpha$ almost surely. Although the objective is similar in spirit, this constraint leads to design issues for the algorithm that are quite different from ours, and are arguably less mathematically elegant. Our setting enforces similar guarantees while bypassing these issues. Finally, Li et al. (2019) consider a problem called combinatorial sleeping bandits, in which the player may pick multiple actions among the K available at every step. The authors also impose that their algorithms satisfy some diversity preserving condition on the choice of the actions, but this condition is only asymptotical.

Note that all these articles refer to "fairness", although we prefer the term "diversity-preserving" to distinguish them from the stream of work discussed below.

Fairness. The framework of individual fairness from Dwork et al. (2012) relies on the idea that "similar individuals should be treated similarly" (there, actions correspond to individuals). This was modeled in stochastic bandits by imposing constraints on the unknown problem, with constraints dictated by the very nature of the problem, hence, being unknown as well. Therefore, the algorithms will need to explore some more in order to learn the constraints while playing the bandit game. The usual tradeoff between exploration and exploitation is therefore modified. We may cite contributions by Joseph et al. (2016), Amani et al. (2019), Liu et al. (2017) and Gillen et al. (2018). This approach is mathematically quite different from the setting considered in this article.

Appendix C. Extended comparison to the linear bandit setting used by Celis et al. (2019) and literature review on structured stochastic bandits

Celis et al. (2019) suggest using linear bandit algorithms in the diversity-preserving setting, unlocking a wide array of methods and regret guarantees. Indeed, they observe that one can discard the knowledge of the true played action A_t and only play using the observation of the rewards Y_t and of the mixed action $\underline{p}_t$. Doing so, the player plays a game where the expected reward associated with the played action $\underline{p}_t$ depends linearly on the said action: this setting is known as linear bandits and is recalled below. An exception to this approach is their CONSTRAINED-ε-GREEDY algorithm, which does use the knowledge of A_t but suffers from two limitations: its regret bound scales as $1/\Delta^2$ instead of $1/\Delta$ and, more importantly, it requires a lower bound on the unknown minimal gap Δ, which makes it quite impractical compared to other knowledge-independent algorithms.

In this appendix, we recall the setting of linear bandits and provide a brief account of the relevant literature on it. We also review the existing literature on other structured bandit settings and bandits with augmented feedback. We then provide intuitions on the limitations of using a reduction to linear bandits for the diversity-preserving bandit game.

C.1. The setting of linear bandits

We refer the interested reader to the monograph by Lattimore and Szepesvári, 2020, Chapter 19 for a longer description. An action set $\mathcal{A} \subset \mathbb{R}^d$ is given to the learner. Some parameter $\boldsymbol{\mu} \in \mathbb{R}^d$ is set but remains unknown to the learner. The latter selects at each step an action $X_t \in \mathcal{A}$ and gets and observes a random reward Y_t such that $\mathbb{E}[Y_t \,|\, X_t] = \langle X_t,\, \boldsymbol{\mu} \rangle$. The expected regret is defined as

$$R_T^{\text{lin}} = T \max_{x \in \mathcal{A}} \langle x,\, \boldsymbol{\mu} \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t,\, \boldsymbol{\mu} \rangle \right].$$

Reduction. The diversity-preserving bandit protocol described in Section 1 can be put in this setting, by having the chosen probability vector $\underline{p}_t$ play the role of X_t. The ambient dimension is then essentially $d = K$. (Notice in particular that the notions of regret coincide.) Therefore, the learner can choose to ignore completely the observation of A_t and use a linear bandit algorithm of her choice, thus transferring the regret guarantees to the diversity-preserving setting. We now discuss the typical regret guarantees achieved in linear bandits.

Standard linear bandit results. A first stream of the literature focuses on generalizations of the UCB algorithm called LinUCB (linear upper confidence bound) or OFUL (optimism in the face of uncertainty for linear bandits); they were introduced by Li et al. (2010) and Chu et al. (2011) and studied by Abbasi-Yadkori et al. (2011). They consider the set $\mathcal{L}$ of bandit models such that the parameter $\boldsymbol{\mu}$ satisfies $\langle x,\, \boldsymbol{\mu} \rangle \in [-1, 1]$ for all $x \in \mathcal{A}$ and the noise $Y_t - \mathbb{E}[Y_t | X_t]$ is sub-Gaussian (with constant less than $1/4$, say). They obtain finite-time distribution-dependent bounds in the case where $\mathcal{A}$ is finite or is a finite polytope; we denote by $\mathcal{A}_{\text{finite}}$ a finite set of points generating $\mathcal{A}$ when it is a finite polytope and let $\mathcal{A}_{\text{finite}} = \mathcal{A}$ when $\mathcal{A}$ is finite. These finite-time distribution-dependent bounds are of the form: there exists a numerical constant C such that for each bandit problem in $\mathcal{L}$,

$$R_T^{\text{lin}} \leqslant C \, \frac{m_2}{\Delta} \big(\ln^2 T + d \ln T + d^2 \ln \ln T \big),$$

where m_2 is a known upper bound on $\|\boldsymbol{\mu}\|_2^2$, and where the gap $\Delta(x)$ of an action $x \in \mathcal{A}$ and the overall gap Δ among suboptimal actions are defined as

$$\Delta(x) = \max_{y \in \mathcal{A}} \langle y - x, \boldsymbol{\mu} \rangle \qquad \text{and} \qquad \Delta = \min\big\{\Delta(x) : x \in \mathcal{A}_{\text{finite}} \text{ s.t. } \Delta(x) > 0\big\}\,.$$

Note that for a fair comparison with our bounds, we should take $m_2 = d = K$, as we assume that $\|\boldsymbol{\mu}\|_\infty \leqslant 1$, which only implies $\|\boldsymbol{\mu}\|_2 \leqslant \sqrt{d}$.

A second stream of the linear bandit literature improves asymptotically on the treatment of the situation where $\mathcal{A}$ is finite or is a polytope and obtains optimal distribution-dependent bounds that only scale with $\ln T$, but at the cost of computational efficiency. Of course, $\ln T$ bounds could have been obtained by playing a plain UCB on $\mathcal{A}_{\text{finite}}$, but they would not involve optimal constant in front of the $\ln T$. The results of Lattimore and Szepesvári (2017), Combes et al. (2017) and Hao et al. (2020) fall in this category.

C.2. Other reductions? Literature review on bandits with augmented feedback.

Actually, linear bandits are a particular case of structured bandits, in which observing the reward associated with an action may provide information about the reward of other actions. This is to be opposed to the vanilla K-armed bandit setting. A lot of recent work (discussed below) has been devoted to general structured bandits, sometimes obtaining bounded regret. Since all these approaches apply to the linear bandit setting, they can also be applied to the diversity-bandit setting. However, each come with some limitations, which may be avoided as the diversity-preserving setting is in fact an easier setting than linear bandits. We may cite the works by Hao et al. (2020), Jun and Zhang (2020), Tirinzoni et al. (2020), and Lattimore and Munos (2014): they all exhibit models (natural exploration in linear contextual bandits and worst-case structures, respectively) in which the introduced algorithms yield bounded regret. However, when applied to the linear bandit problems that emerge from the linear bandit setting, these approaches cannot give bounded regret.

Indeed, and this is the fundamental caveat of applying linear bandit methods, when neglecting the knowledge of A_t, the problem faced by the player becomes *exactly* a linear bandit problem. Therefore, these methods are subject to the linear bandit lower bounds. In particular, for typical (fixed) finite action sets, the lower bound of Lattimore and Szepesvári (2017) implies that the regret incurred by linear bandit algorithms must grow logarithmically as $T \to \infty$ on any problem. By contrast, we show that our approach which, uses the knowledge of A_t, can yield finite regret. Note also that Caron et al. (2012) and Degenne et al. (2018) consider K-armed bandit models with some extra feedback and provide bounded regret guarantees then. In the rest of this section, we provide some more insights and intuitions on the nature of the improvements obtained when taking this extra piece of information into account.

C.3. Intuitions on why taking A_t into account helps

We provide two intuitions, one linked to lower bounds on the regret and the other one linked to the upper bounds on the regret.

As far as lower bounds are concerned. As is clear from the proofs in Section 3, see, e.g., Equation (11), lower bounds rely on the ability to discriminate between two bandit problems $\underline{\nu}$ and $\underline{\nu}'$. Under the problem $\underline{\nu}$ and conditionally to the choice of a distribution $\underline{p}_t$ over the arms, the learner

sees the payoff Y_t as distributed according to some unconditional distribution when A_t is not taken into account, and the conditional distribution ν_{A_t} when A_t is taken into account:

$$Y_t \sim \sum_{a\in[K]} p_{t,a}\nu_a \qquad \text{and} \qquad Y_t \,|\, A_t \sim \nu_{A_t}\,,$$

respectively. Conditionally to the choice of $\underline{p}_t$, the Kullback-Leibler divergences between the distributions of Y_t under $\underline{\nu}$ and $\underline{\nu}'$ are therefore given by

$$\underbrace{\mathrm{KL}\left(\sum_{a\in[K]} p_{t,a}\nu_a,\ \sum_{a\in[K]} p_{t,a}\nu'_a\right)}_{\text{without } A_t} \leqslant \underbrace{\mathbb{E}\big[\mathrm{KL}(\nu_{A_t},\nu'_{A_t})\big]}_{\text{with } A_t} = \sum_{a\in[K]} p_{t,a}\,\mathrm{KL}(\nu_a,\nu'_a)\,,$$

where the inequality is by convexity of KL. We of course prefer the larger quantity to derive the largest possible limiting constant.

As far as upper bounds are concerned. Here, we discuss more concretely how the knowledge of A_t can improve our algorithms: using A_t helps building tighter confidence bounds on $\underline{\mu}$. As a reminder, the confidence region considered in Section 2 is the hyper-rectangle

$$\left\{(\mu_1,\dots,\mu_K)\in\mathbb{R}^K \ :\ \forall a\in[K],\quad |\widehat{\mu}_a(t)-\mu_a| \leqslant \sqrt{\frac{2\ln t}{\max(N_a(t),1)}}\right\} \tag{16}$$

obtained from the Hoeffding–Azuma inequality and by treating each coordinate separately. Now, LinUCB for linear bandits (see references above) rather relies on an ellipsoid of confidence, constructed as indicated below in (17). Figure 1 compares the confidence regions (16) and (17) on some simulated data.

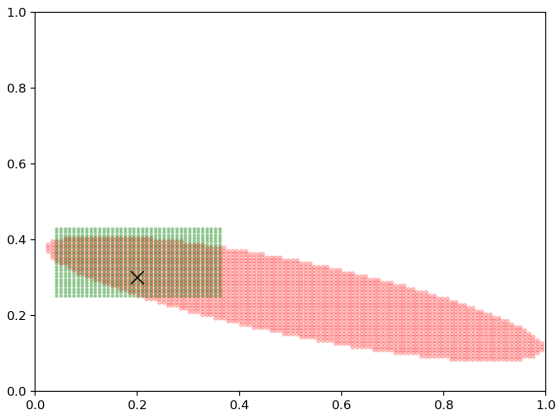


Figure 1: Comparison of confidence sets for Bernoulli observations generated from the three probability vectors $\underline{p}_1 = (0.1, 0.9)$, $\underline{p}_2 = (0.2, 0.8)$, $\underline{p}_3 = (0.4, 0.6)$ and true mean vector $(\mu_1,\mu_2) = (0.2, 0.3)$. Each $\underline{p}_i$, for $i \in \{1, 2, 3\}$, was selected 100 times to draw actions $A_t \in \{1, 2\}$, after which rewards $Y_t \sim \mathrm{Ber}(\mu_{A_t})$ were generated; this thus provided $T = 300$ observations. The true mean vector is shown as a cross. The red area depicts the ellipsoid defined in (17) without the knowledge of the A_t, whereas the green rectangle is the one from (16) and relies on the A_t.

We denote by $\mathbb{X}_t$ the matrix whose rows are the mixed actions used over time, $\underline{p}_1^\top, \dots, \underline{p}_t^\top$, and let $Y_{1:t} = (Y_1, \dots, Y_t)^\top$ be the column vector of rewards obtained. Then, for a parameter $\lambda > 0$, we define

$$V_t^\lambda = \lambda \mathrm{I}_d + \mathbb{X}_t^\top \mathbb{X}_t \qquad \text{and} \qquad \underline{\widehat{\mu}}_t^{\text{lin}} = \left(V_t^\lambda\right)^{-1} \mathbb{X}_t^\top Y_{1:t}\,.$$

The ellipsoid considered is then of the form

$$\left\{ \underline{\mu} \in \mathbb{R}^d \ : \ \sqrt{\left\langle \underline{\mu} - \underline{\widehat{\mu}}_t^{\text{lin}},\, V_t^\lambda \big(\underline{\mu} - \underline{\widehat{\mu}}_t^{\text{lin}}\big) \right\rangle} \leqslant f_{\sqrt{\ln}}(t) + \lambda\sqrt{d} \right\}, \tag{17}$$

where $f_{\sqrt{\ln}}(t)$ is some closed-form function of order $\sqrt{\ln t}$.

Appendix D. Full proof of the upper bounds

For the common part of the proofs of Cases 2 and 3, the following lemma was used. It is standard in the literature of vanilla K–armed bandits: we simply note that it also holds in our setting. As we have no direct control on arms pulled, we cannot ensure in a deterministic manner that $N_a(t) \geqslant 1$ a.s., hence we take the maximum between $N_a(t)$ and 1.

Lemma 4 *Consider a bandit problem $\underline{\nu}$ in $\mathcal{D}_{[0,1]}$. For $t \geqslant 2$, if the actions $A_1, \dots, A_t$ and rewards $Y_1, \dots, Y_t$ were generated according to Protocol 1, then*

$$\mathbb{P}\big(\mathcal{E}(t)\big) = \mathbb{P}\left\{\text{For all } a \in [K], \quad \big|\mu_a - \widehat{\mu}_a(t)\big| \leqslant \sqrt{\frac{2\ln t}{\max\big\{N_a(t), 1\big\}}}\right\} \geqslant 1 - 2Kt^{-3}\,.$$

Proof By optional skipping[2] (see Theorem 5.2 of Doob, 1953, Chapter III, p. 145, see also Chow and Teicher, 1988, Section 5.3), we can replace the random quantities depending on the observations from a fixed arm by their i.i.d. analogue. More precisely, for each arm $a \in [K]$, by defining $\widehat{\mu}_{a,n}$ as an empirical average of n i.i.d. random variables with distribution ν_a, we have

$$\begin{aligned}
&\mathbb{P}\left\{\big|\mu_a - \widehat{\mu}_a(t)\big| \geqslant \sqrt{\frac{2\ln t}{\max\big\{N_a(t), 1\big\}}}\right\} \\
&\leqslant \mathbb{P}\left\{\exists n \in \{0, 1, \dots, t\} \;:\; \big|\mu_a - \widehat{\mu}_{a,n}\big| \geqslant \sqrt{\frac{2\ln t}{\max\{n, 1\}}}\right\} \\
&\leqslant \sum_{n=1}^{t} \mathbb{P}\left\{\big|\mu_a - \widehat{\mu}_{a,n}\big| \geqslant \sqrt{\frac{2\ln t}{\max\{n, 1\}}}\right\} \leqslant \sum_{n=1}^{t} 2t^{-4} = 2t^{-3}\,,
\end{aligned}$$

where the case $n = 0$ was dropped for reasons explained below, where the second inequality follows from a union bound and the third inequality, from Hoeffding's inequality. Note indeed that $n = 0$ and $t \geqslant 2$ are incompatible given the event considered: we defined $\widehat{\mu}_{a,0}$ to be 1; when $n = 0$, the event considered amounts to $|1 - \mu_a| \geqslant \sqrt{2\ln t}$, where $\sqrt{2\ln t}$ is larger than $\sqrt{2\ln 2} > 1$ when $t \geqslant 2$. The claimed inequality follows from a final union bound over $a \in [K]$. ∎

D.1. Case 3: Proof of the upper bound (6) on t_0

We actually also prove a lower bound on t_0:

$$\frac{16}{\Delta^2\ell} - 1 \leqslant t_0 \leqslant \frac{32}{\Delta^2\ell}\ln\left(\frac{16}{\Delta^2\ell}\right).$$

This result is a special case of the following more general result, with $a = \Delta^2\ell/16 < \ln(2)/2$.

Lemma 5 *Let $a \in \big(0, \ln(2)/2\big)$. Define*

$$x_0 = \sup\big\{x \in (0, +\infty) : \ln x > ax\big\} \qquad \textit{and} \qquad n_0 = \max\big\{n \in \{1, 2, 3, \dots\} : \ln n > an\big\}\,.$$

Then $x_0 - 1 \leqslant n_0 \leqslant x_0$ and

$$\frac{1}{a} < x_0 < \frac{2}{a}\ln\left(\frac{1}{a}\right).$$

2. Sometimes called optional sampling.

Proof First note that n_0 and x_0 are well defined since $\ln(2) > 2a$ (by assumption) and the function $\psi : x \mapsto ax - \ln x$ satisfies $\psi(x) \to +\infty$ as $x \to +\infty$. Note also that ψ is continuous, decreasing on $(0, 1/a]$, and increasing on $[1/a, +\infty)$.

The inequality $x_0 > 1/a$ follows from $\psi(1/a) = 1 + \ln a < 0$ (since $a < \ln(2)/2 < 1/\mathrm{e}$) and by continuity of ψ.

For the stated upper bound on x_0, we first note that

$$\psi\left(\frac{2}{a}\ln\frac{1}{a}\right) = a\,\frac{2}{a}\ln\frac{1}{a} - \ln\left(\frac{2}{a}\ln\frac{1}{a}\right) = 2\ln\frac{1}{a} - \ln\frac{1}{a} - \underbrace{\ln\left(2\ln\frac{1}{a}\right)}_{<\ln(1/a)} > 0\,,$$

where we used that $2\ln u < u$ for all $u > 0$. Given that $a < 1$ and thus that $(2/a)\ln(1/a) > 1/a$, and since ψ is continuous and increasing on $[1/a, +\infty)$, the inequality above indeed shows that $x_0 < (2/a)\ln(1/a)$.

The inequality $n_0 \leqslant x_0$ is straightforward by definitions. We now prove that $x_0 - 1 \leqslant n_0$. By definition, n_0 is the largest integer $n \geqslant 1$ such that $\psi(n) < 0$, so that $\psi(n_0) < 0$ while $\psi(n_0 + 1) \geqslant 0$. Given the variations of ψ, we therefore have that $n_0 + 1$ is on the increasing branch of ψ, that is, $n_0 + 1 \geqslant 1/a$. Again since ψ increases on $[1/a, +\infty)$ and given that $x_0 > 1/a$, as well as $\psi(x_0) = 0$ by continuity of ψ, we proved that $x_0 \leqslant n_0 + 1$. ■

D.2. Case 2: Proof of the finite sum (8)

We will actually prove a stronger result, namely

$$\mathbb{P}\left\{N_\star(t) \leqslant \frac{t}{2}\right\} = \mathcal{O}(1/t^2)\,.$$

We first tie the number of optimal pulls $N_\star(t)$ to the (non-expected) instantaneous regret: by definition of the minimal gap Δ,

$$\sum_{s=1}^{t}\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle \geqslant \Delta\sum_{s=1}^{t}\mathbb{1}_{\{\underline{p}_s \notin \mathrm{Opt}(\underline{\nu})\}} = \Delta\big(t - N_\star(t)\big)\,, \quad \text{thus} \quad N_\star(t) \geqslant t - \frac{1}{\Delta}\sum_{s=1}^{t}\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle\,,$$

where $\underline{p}^\star$ denotes any optimal mixed action. Therefore, we only need to prove that

$$\mathbb{P}\left\{\sum_{s=1}^{t}\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle \geqslant \frac{\Delta t}{2}\right\} = \mathcal{O}(1/t^2)\,. \tag{18}$$

This is a weak high-probability bound on the non-expected cumulative regret: we only want to show that it grows, with high probability, slower than the linear quantity $\Delta t/2$. We consider to that end the following lemma (which will also be used for the proof of Case 1).

Lemma 6 *For $s \geqslant 2$, under the event $\mathcal{E}(s-1)$ defined in (4),*

$$\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle \leqslant 2\sum_{a\in[K]} p_{s,a}\sqrt{\frac{2\ln(s-1)}{\max\big\{N_a(s-1), 1\big\}}}\,.$$

Proof By definition of the algorithm, we have $\langle \underline{p}^\star - \underline{p}_s,\, \underline{U}(s-1)\rangle \leqslant 0$, so that

$$\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle = \langle \underline{p}^\star,\, \underline{\mu} - \underline{U}(s-1)\rangle + \underbrace{\langle \underline{p}^\star - \underline{p}_s,\, \underline{U}(s-1)\rangle}_{\leqslant 0} + \langle \underline{p}_s,\, \underline{U}(s-1) - \underline{\mu}\rangle\,.$$

Now, under $\mathcal{E}(s-1)$,

$$\begin{aligned}\forall a \in [K], \qquad 0 \leqslant U_a(s-1) - \mu_a &= \widehat{\mu}_a(s-1) + \sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}} - \mu_a \\ &\leqslant 2\sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}}\,.\end{aligned}$$

Therefore, under $\mathcal{E}(s-1)$, substituting in the first bound of this proof, we get

$$\begin{aligned}\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle &\leqslant \langle \underline{p}^\star,\, \underline{\mu} - \underline{U}(s-1)\rangle + \langle \underline{p}_s,\, \underline{U}(s-1) - \underline{\mu}\rangle \\ &\leqslant 0 + 2\sum_{a\in[K]} p_{s,a}\sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}}\,.\end{aligned}$$

■

Recall the aim (18). Since the events $\overline{\mathcal{E}(s-1)}$ in Lemma 6 might have a non-negligible probability for small values of s, we leave out the first $\tau_t := \lfloor \Delta t/4 \rfloor$ rounds (note that $\tau_t \geqslant 1$ for $t \geqslant 4/\Delta$). More precisely, in order to prove (18), noting that $\langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle$ lies in $[-1,1]$ as $\underline{\nu}$ lies in $\mathcal{D}_{[0,1]}$, it suffices to show that

$$\mathbb{P}\left\{\sum_{s=\tau_t}^{t} \langle \underline{p}^\star - \underline{p}_s,\, \underline{\mu}\rangle \geqslant \frac{\Delta t}{4}\right\} = \mathcal{O}(1/t^2)\,.$$

By Lemma 6, and noting that $\tau_t \geqslant 2$ for $t \geqslant 8/\Delta$, we thus only need to prove that

$$\mathbb{P}\left(\bigcup_{s=\tau_t}^{t} \overline{\mathcal{E}(s-1)}\right) + \mathbb{P}\left\{2\sum_{s=\tau_t}^{t}\sum_{a\in[K]} p_{s,a}\sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}} \geqslant \frac{\Delta t}{4}\right\} = \mathcal{O}(1/t^2)\,.$$

By a union bound and given the fact that $\mathbb{P}\big(\overline{\mathcal{E}(s)}\big) \leqslant 2Ks^{-3}$ by Lemma 4, the first probability above is at most of $K(\tau_t - 1)^{-2} = \mathcal{O}(1/t^2)$. Therefore, it suffices to prove that:[3]

$$\mathbb{P}\left\{2\sum_{s=2}^{t}\sum_{a\in[K]} p_{s,a}\sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}} \geqslant \frac{\Delta t}{4}\right\} = \mathcal{O}(1/t^2)\,. \tag{19}$$

We apply the Hoeffding-Azuma inequality (see, e.g., Lemma A.7 in Cesa-Bianchi and Lugosi 2006) to the martingale

$$M_t \overset{\text{def}}{=} 2\sum_{s=2}^{t}\sum_{a\in[K]} \big(p_{s,a} - \mathbb{1}_{\{A_s=a\}}\big)\sqrt{\frac{2\ln(s-1)}{\max\{N_a(s-1),1\}}}$$

3. Adding the (nonnegative) terms from $s=2$ to $s=\tau_t - 1$ in the sum makes the goal only harder.

with (predictable) increments lying in a range of width smaller than $2\sqrt{2\ln t}$ and by picking a risk level of $\delta_t = t^{-4}$; we obtain:

$$\mathbb{P}\Big\{ M_t \geqslant 2\sqrt{2\ln t}\,\sqrt{2t\ln t} \Big\} \leqslant t^{-4}\,. \tag{20}$$

Now, for each $a \in [K]$, as $N_a(s-1)$ only increases (by 1) when $A_s = a$, and since $N_a(t-1) \leqslant t-1$, we have the deterministic (crude) bound

$$\begin{aligned} C_t \overset{\text{def}}{=} 2\sum_{s=2}^{t} \mathbb{1}_{\{A_s=a\}} \sqrt{\frac{2\ln(s-1)}{\max\big\{N_a(s-1),1\big\}}} &\leqslant 2\sqrt{2\ln t}\sum_{s=2}^{t} \frac{\mathbb{1}_{\{A_s=a\}}}{\sqrt{\max\big\{N_a(s-1),1\big\}}} \\ &\leqslant 2\sqrt{2\ln t}\sum_{n=0}^{t-1} \frac{1}{\sqrt{\max\{n,1\}}} \leqslant 2\sqrt{2\ln t}\big(1+2\sqrt{t-1}\big)\,. \end{aligned} \tag{21}$$

All in all, with the processes M_t and C_t introduced above,

$$\left\{ 2\sum_{s=2}^{t}\sum_{a\in[K]} p_{s,a}\sqrt{\frac{2\ln(s-1)}{\max\big\{N_a(s-1),1\big\}}} \geqslant \frac{\Delta t}{4} \right\} = \left\{ C_t + M_t \geqslant \frac{\Delta t}{4} \right\}.$$

The sum

$$2\sqrt{2\ln t}\,\sqrt{2t\ln t} + 2\sqrt{2\ln t}\big(1+2\sqrt{t-1}\big) = \mathcal{O}\big(\sqrt{t}\ln t\big)$$

of the high-probability bound exhibited on M_t and of the deterministic bound exhibited on C_t is smaller than $\Delta t/4$ after some step t_0. Hence, for $t \geqslant t_0$, the value $\Delta t/4$ cannot be achieved unless the high-probability bound on M_t does not hold:

$$\left\{ C_t + M_t \geqslant \frac{\Delta t}{4} \right\} \subseteq \left\{ M_t \geqslant 2\sqrt{2\ln t}\,\sqrt{2t\ln t} \right\}.$$

This remark, together with the deviation bound (20), proves that the sufficient aim (19) is true. This concludes the proof of Case 2.

D.3. Case 1: Complete proof

The beginning of this proof follows a standard proof scheme for linear bandits. For $t \geqslant 1$, we denote by

$$r_t = \big\langle \underline{p}^\star - \underline{p}_t\,, \underline{\mu} \big\rangle$$

the (non-expected) instantaneous regret; it was already considered in Lemma 6. The regret to be controlled corresponds to

$$R_T(\underline{\nu}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t \right].$$

We requested in the definition of Algorithm 1 that mixed actions be selected only in $\mathrm{Ext}(\mathcal{P})$. Doing so, the (non-expected) instantaneous regret suffered from playing $\underline{p}_t \in \mathrm{Ext}(\mathcal{P})$ is either 0, if $\underline{p}_t$ is optimal, or at least Δ if $\underline{p}_t$ is suboptimal. This simple observation leads to the crude upper bound

$$r_t \leqslant \frac{r_t^2}{\Delta}\,.$$

Now, under $\mathcal{E}(t-1)$, Lemma 6 followed by an application of the Cauchy-Schwarz inequality yield

$$\begin{aligned} r_t^2 &\leqslant \left(2\sum_{a\in[K]} p_{t,a}\sqrt{\frac{2\ln(t-1)}{\max\{N_a(t-1),1\}}}\right)^2 \\ &\leqslant 8\left(\sum_{a\in[K]}\frac{p_{t,a}}{\max\{N_a(t-1),1\}}\right)\left(\sum_{a\in[K]} p_{t,a}\ln(t-1)\right) \\ &\leqslant 8\left(\sum_{a\in[K]}\frac{p_{t,a}}{\max\{N_a(t-1),1\}}\right)\ln T\,. \end{aligned}$$

Since A_t is drawn at random given $\underline{p}_t$, which is determined by the information available at the beginning of round t, just as $N_a(t-1)$ is, the tower rule indicates that for all $t \geqslant 1$ and $a \in [K]$,

$$\mathbb{E}\left[\frac{p_{t,a}}{\max\{N_a(t-1),1\}}\right] = \mathbb{E}\left[\frac{\mathbb{1}_{\{A_t=a\}}}{\max\{N_a(t-1),1\}}\right].$$

Taking into account the fact that when $\mathcal{E}(t-1)$ is not satisfied, we have $r_t^2 \leqslant 1$ as $\underline{\nu}$ lies in $\mathcal{D}_{[0,1]}$, we proved so far

$$\mathbb{E}\big[r_t^2\big] \leqslant \mathbb{P}\big(\mathcal{E}(t-1)\big) + 8\sum_{a\in[K]}\mathbb{E}\left[\frac{\mathbb{1}_{\{A_t=a\}}}{\max\{N_a(t-1),1\}}\right]\ln T.$$

Lemma 4 shows that $\mathbb{P}\big(\mathcal{E}(t-1)\big) \leqslant 2K(t-1)^{-3}$ for $t \geqslant 3$ (and we resort to the trivial bound 1 for $t = 1$ or $t = 2$).

Now come the final steps of the proof: they are specific to our diversity-preserving setting and short-cut the classical proof scheme. The same kind of deterministic argument as used in (21) shows that for each $a \in [K]$,

$$\begin{aligned} \sum_{t=1}^{T}\frac{\mathbb{1}_{\{A_t=a\}}}{\max\{N_a(t-1),1\}} = \sum_{n=0}^{N_a(T-1)}\frac{1}{\max\{n,1\}} &\leqslant 2+\ln\Big(\max\big\{N_a(T-1),1\big\}\Big) \\ &\leqslant 2+\ln\big(1+N_a(T-1)\big)\,, \end{aligned}$$

where we used that $1 + 1/2 + \ldots + 1/N \leqslant 1 + \ln N$ for $N \geqslant 1$. Collecting all bounds together, we proved

$$\sum_{t=1}^{T}\mathbb{E}\big[r_t^2\big] \leqslant 2 + 2K\underbrace{\sum_{t=3}^{T}(t-1)^{-3}}_{<1/2} + 8\sum_{a\in[K]}\Big(2+\mathbb{E}\Big[\ln\big(1+N_a(T-1)\big)\Big]\Big)\ln T\,.$$

By concavity of the logarithm,

$$\frac{1}{K}\sum_{a\in[K]}\ln\big(1+N_a(T-1)\big) \leqslant \ln\Bigg(1+\frac{1}{K}\underbrace{\sum_{a\in[K]}N_a(T-1)}_{=T-1}\Bigg) \leqslant \ln(1+T/K)\,.$$

We summarize all calculations performed so far into:

$$R_T(\underline{\nu}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \leqslant \frac{1}{\Delta}\sum_{t=1}^{T}\mathbb{E}\big[r_t^2\big] \leqslant \frac{1}{\Delta}\big(2 + K + 16K\ln T + 8K\ln(1+T/K)\ln T\big)\,.$$

The final bound is transformed into the stated $\Big(2+K+24K\big(\ln(1+T)\big)^2\Big)/\Delta$ for better readability.

Appendix E. Full proof of the lower bound

E.1. Proof of the reduction

We first note that in Protocol 1, whether a strategy picks a deterministic mixed action $\underline{p}_t$ (based on the past) or a distribution ρ_t over mixed actions in $\mathcal{P}$ is irrelevant, as the strategy needs to draw an arm A_t and only observes the payoff obtained by picking A_t. In the first case, A_t is drawn in a one-step randomization, while in the second case, A_t is drawn in a two-step randomization which is equivalent to picking the deterministic mixed action

$$\underline{p}_t = \int_{\mathcal{P}} \underline{p} \, \mathrm{d}\rho_t(\underline{p}) \, .$$

When a strategy picks mixed actions in $\mathcal{P}$, whether it is deterministic or randomized is irrelevant.

Based on the same intuition, we now show that it suffices to restrict one's attention to strategies playing in the finite set $\mathrm{Ext}(\mathcal{P})$ generating $\mathcal{P}$. Any mixed action $\underline{q} = \underline{p}_t$ may be decomposed as a convex combination of elements of $\mathrm{Ext}(\mathcal{P})$. We may define a function

$$\Phi : \underline{q} \in \mathcal{P} \longmapsto \Phi(\underline{q}) = \big(\Phi_{\underline{p}}(\underline{q})\big)_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} \in [0,1]^{\mathrm{Ext}(\mathcal{P})}$$

such that all images $\Phi(\underline{q})$ are actually convex weights and

$$\forall \underline{q} \in \mathcal{P}, \qquad \underline{q} = \sum_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} \Phi_{\underline{p}}(\underline{q}) \, \underline{p} \, .$$

We may interpret convex weights $\Phi(\underline{q})$ as probability distributions over $\mathrm{Ext}(\mathcal{P})$. With this in mind, we note that a (deterministic) strategy ψ picking mixed actions $\underline{p}_t$ (based on the information available: past actions A_s and rewards Y_s, with $s \leqslant t-1$) gets the same expected payoffs as the (randomized) strategy ψ_Φ that first picks a mixed action $\underline{P}_t$ in $\mathrm{Ext}(\mathcal{P})$ at random according to $\Phi(\underline{p}_t)$, and then draws the action A_t according to $\underline{P}_t$. This, again, holds because only the choice of the pure action A_t matters. In particular, if the strategy ψ is UFC over $\mathcal{D}$ given $\mathcal{P}$, then so is ψ_Φ.

The final issue to clarify is that the proof of Theorem 2 following this reduction holds for deterministic and randomized strategies ψ_Φ (again because only the actions A_t drawn matter).

E.2. Rest of the proof of the first part of Theorem 2

This proof scheme is standard and was introduced by Graves and Lai (1997).

Part A: The divergence $\mathcal{I}_T$ is larger than $\ln T$ in the limit. We prove that $\mathcal{I}_T = \mathrm{KL}\left(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\right)$ is asymptotically larger than $\ln T$.

Lemma 7 *For all models $\mathcal{D}$ with finite first moments, for all diversity-preserving sets $\mathcal{P}$, for all strategies that are uniformly fast convergent [UFC] over $\mathcal{D}$ given $\mathcal{P}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal action $\underline{p}^\star(\underline{\nu})$, for all bandit problems $\underline{\nu}'$ in $\mathcal{D}$ for which $\underline{p}^\star(\underline{\nu})$ is suboptimal,*

$$\liminf_{T \to \infty} \frac{\mathrm{KL}\left(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\right)}{\ln T} \geqslant 1 \, .$$

Proof We denote by $\mathrm{kl}(p,q) = p\ln(p/q) + (1-p)\ln\big((1-p)/(1-q)\big)$ the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q. By the data-processing inequality for $[0,1]$–valued random variables (see Section 2.1 in Garivier et al., 2019) and by the standard inequality $\mathrm{kl}(p,q) \geqslant p\ln(1/q) - \ln 2$,

$$\begin{aligned}\mathrm{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big) &\geqslant \mathrm{kl}\left(\mathbb{E}_{\underline{\nu}}\left[\frac{N_{\underline{p}^\star(\underline{\nu})}(T)}{T}\right], \mathbb{E}_{\underline{\nu}'}\left[\frac{N_{\underline{p}^\star(\underline{\nu})}(T)}{T}\right]\right) \\ &\geqslant \mathbb{E}_{\underline{\nu}}\left[\frac{N_{\underline{p}^\star(\underline{\nu})}(T)}{T}\right]\ln\left(\frac{T}{\mathbb{E}_{\underline{\nu}'}\Big[N_{\underline{p}^\star(\underline{\nu})}(T)\Big]}\right) - \ln 2\,.\end{aligned}$$

The strategy is UFC, we therefore have $R_T(\underline{\nu})/T \to 0$. Given the decomposition (3) of the regret and given that $\Delta(\underline{\nu}) > 0$ for all $\underline{\nu} \in \mathrm{Ext}(\mathcal{P})$ with $\underline{\nu} \neq \underline{p}^\star(\underline{\nu})$ since $\underline{p}^\star(\underline{\nu})$ is the only optimal action for $\underline{\nu}$, we have, on the one hand,

$$\mathbb{E}_{\underline{\nu}}\left[\frac{N_{\underline{p}^\star(\underline{\nu})}(T)}{T}\right] \xrightarrow[T\to\infty]{} 1\,.$$

On the other hand, given that $\underline{p}^\star(\underline{\nu})$ is suboptimal for $\underline{\nu}'$, the suboptimality gap $\Delta_{\underline{\nu}'}\big(\underline{p}^\star(\underline{\nu})\big)$ of $\underline{p}^\star(\underline{\nu})$ in the bandit problem $\underline{\nu}'$ is positive and

$$R_T(\underline{\nu}') \geqslant \Delta_{\underline{\nu}'}\big(\underline{p}^\star(\underline{\nu})\big)\,\mathbb{E}_{\underline{\nu}'}\Big[N_{\underline{p}^\star(\underline{\nu})}(T)\Big]\,, \qquad \text{or equivalently,} \qquad \mathbb{E}_{\underline{\nu}'}\Big[N_{\underline{p}^\star(\underline{\nu})}(T)\Big] \leqslant \frac{R_T(\underline{\nu}')}{\Delta_{\underline{\nu}'}\big(\underline{p}^\star(\underline{\nu})\big)}\,.$$

The strategy being UFC, we have $R_T(\underline{\nu}') = o(T^\alpha)$ for all $\alpha > 0$; in particular, for all $\varepsilon > 0$, for all T large enough, $\mathbb{E}_{\underline{\nu}'}\Big[N_{\underline{p}^\star(\underline{\nu})}(T)\Big] \leqslant T^\varepsilon$. Putting all inequalities together, we proved

$$\frac{\mathrm{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big)}{\ln T} \geqslant \underbrace{\mathbb{E}_{\underline{\nu}}\left[\frac{N_{\underline{p}^\star(\underline{\nu})}(T)}{T}\right]}_{\longrightarrow 1}\ln\Bigg(\underbrace{\frac{T}{\mathbb{E}_{\underline{\nu}'}\Big[N_{\underline{p}^\star(\underline{\nu})}(T)\Big]}}_{\geqslant T^{1-\varepsilon}}\Bigg)\frac{1}{\ln T} - \underbrace{\frac{\ln 2}{\ln T}}_{\longrightarrow 0}\,.$$

Therefore, as $T \to \infty$, we have

$$\liminf_{T\to\infty} \frac{\mathrm{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big)}{\ln T} \geqslant 1 - \varepsilon\,,$$

and the claimed result follows by taking $\varepsilon \to 0$. ■

Part B: Considering cluster points. We conclude the proof of the first part of Theorem 2. Let c be a cluster point of the sequence $R_T/\ln T$. If $c = +\infty$ is the only value, then $R_T/\ln T \to +\infty$ and the result is proved. Otherwise, take a finite c. We denote by $(T_m)_{m\geqslant 1}$ an increasing sequence such that $R_{T_m}/\ln T_m \to c$. In view of the decomposition (10) and since $\Delta(\underline{p}) > 0$ and $n_{T_m}(\underline{p}) \geqslant 0$ for all $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ with $\underline{p} \neq \underline{p}^\star(\underline{\nu})$, these sequences $n_{T_m}(\underline{p})$ are bounded. Hence, we may extract a subsequence $(T_{m_k})_{k\geqslant 1}$ from (T_m) such that all sequences $n_{T_{m_k}}(\underline{p})$ converge as $k \to \infty$, to limits

denoted by $n(\underline{p})$. This only holds for $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ with $\underline{p} \neq \underline{p}^\star(\underline{\nu})$. The final component $n\big(\underline{p}^\star(\underline{\nu})\big)$ is defined in some arbitrary way (its value will be irrelevant).

Now, given $\underline{\nu}' \in \mathrm{ALT}(\underline{\nu})$, the quantity $\mathcal{I}_T$ defined in (11) and rewritten in (12) satisfies, in view of Lemma 7:

$$\liminf_{T\to+\infty} \sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} n_T(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a,\nu'_a) = \liminf_{T\to+\infty} \frac{\mathcal{I}_T}{\ln T} \geqslant 1\,.$$

Note that because of the inner summation over the a such that $p_a^\star(\underline{\nu}) = 0$, the summation in the left-most term can be restricted to $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ with $\underline{p} \neq \underline{p}^\star(\underline{\nu})$:

$$\sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} n_T(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a,\nu'_a) = \sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n_T(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a,\nu'_a)\,.$$

All in all, considering rather the subsubsequence $(T_{m_k})_{k\geqslant 1}$, we get, by identity of the $\liminf$,

$$\sum_{\substack{\underline{p}\in\mathrm{Ext}(\mathcal{P})\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a,\nu'_a) \geqslant 1\,.$$

This holds for all $\underline{\nu}' \in \mathrm{ALT}(\underline{\nu})$ and shows that the vector $\big(n(\underline{p})\big)_{\underline{p}\in\mathrm{Ext}(\mathcal{P})}$ thus satisfies the constraints (13).

In conclusion and in view of the decomposition (10), we have just shown that all finite cluster points c of $R_T/\ln T$ are of the form

$$\sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} \Delta(\underline{p})\, n(\underline{p}) \qquad \text{for some } \big(n(\underline{p})\big)_{\underline{p}\in\mathrm{Ext}(\mathcal{P})} \in \mathbb{R}_+^{\mathrm{Ext}(\mathcal{P})} \text{satisfying the constraints (13).}$$

Note that the value for $n(\underline{p}^\star)$ is irrelevant as $\Delta\big(\underline{p}^\star(\underline{\nu})\big) = 0$. Hence, the $\liminf$ is in particular larger than or equal to the infimum over these quantities, which is exactly the quantity $c\big(\mathrm{Ext}(\mathcal{P}),\underline{\nu}\big)$ defined in Definition 3.

E.3. Proof that Assumption 3.1 holds true when $\mathcal{D}$ is convex

In this section we assume that $\mathcal{D}$ is a convex subset of the set of all probability distributions on the real line, and show that Assumption 3.1 holds true.

Let $\underline{\nu}$ be a bandit problem in $\mathcal{D}$, assume that the set $\mathrm{ALT}(\underline{\nu})$ is non-empty, and consider any $\underline{\nu}' \in \mathrm{ALT}(\underline{\nu})$. Next we show that for $\lambda \in (0,1)$ sufficiently small, the modified problem

$$\underline{\nu}'' = (1-\lambda)\underline{\nu}' + \lambda\underline{\nu} \tag{22}$$

(which still belongs to $\mathcal{D}$ by convexity of $\mathcal{D}$) meets the requirements of Assumption 3.1. Indeed, first note that for all $\lambda \in (0,1)$ and all $a \in [K]$, the Radon-Nikodym derivative $\mathrm{d}\nu_a/\mathrm{d}\nu''_a$ is bounded by $1/\lambda$, so that

$$\mathrm{KL}(\nu_a,\nu''_a) \leqslant \ln(1/\lambda) < +\infty\,.$$

To conclude, it suffices to show that $\underline{\nu}'' \in \textsc{Alt}(\underline{\nu})$. First, for all actions $a \in [K]$ such that $p^\star_a(\underline{\nu}) > 0$, we have $\nu_a = \nu'_a = \nu''_a$ since $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$ and given the definition (22). Second, $p^\star(\underline{\nu}) \notin \mathrm{Opt}(\underline{\nu}'')$ for λ small enough: as $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$, we have by definition $p^\star(\underline{\nu}) \notin \mathrm{Opt}(\underline{\nu}')$, which means that there exists $\underline{p} \in \mathcal{P}$ such that

$$\big\langle \underline{p}, \underline{\mu}' \big\rangle > \big\langle p^\star(\underline{\nu}), \underline{\mu}' \big\rangle \;, \qquad \text{while} \qquad \big\langle \underline{p}, \underline{\mu} \big\rangle \leqslant \big\langle p^\star(\underline{\nu}), \underline{\mu} \big\rangle \;.$$

However, taking $\lambda \in (0,1)$ small enough will still guarantee $\big\langle \underline{p}, \underline{\mu}'' \big\rangle > \big\langle p^\star(\underline{\nu}), \underline{\mu}'' \big\rangle$ for λ. We thus showed that $\underline{\nu}'' \in \textsc{Alt}(\underline{\nu})$ for $\lambda > 0$ small enough. All the above entails that $\underline{\nu}''$ meets the requirements of Assumption 3.1, which concludes the proof.